Journal of Biomedical Informatics 56 (2015) 57–64

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Comparing image search behaviour in the ARRS GoldMiner search engine and a clinical PACS/RIS

Maria De-Arteaga [a,*], Ivan Eggel [b], Bao Do [d], Daniel Rubin [d], Charles E. Kahn Jr. [e], Henning Müller [b,c]

[a] Carnegie Mellon University, Pittsburgh, PA, USA
[b] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[c] Department of Radiology, University Hospitals and University of Geneva, Switzerland
[d] Department of Radiology, School of Medicine, Stanford University, CA, USA
[e] Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 9 October 2014
Revised 17 April 2015
Accepted 22 April 2015
Available online 19 May 2015

Keywords:
Medical image search
Log file analysis
User behaviour
Information retrieval

ABSTRACT

Information search has changed the way we manage knowledge and the ubiquity of information access has made search a frequent activity, whether via Internet search engines or increasingly via mobile devices. Medical information search is in this respect no different and much research has been devoted to analyzing the way in which physicians aim to access information. Medical image search is a much smaller domain but has gained much attention as it has different characteristics than search for text documents. While web search log files have been analysed many times to better understand user behaviour, the log files of hospital internal systems for search in a PACS/RIS (Picture Archival and Communication System, Radiology Information System) have rarely been analysed. Such a comparison between a hospital PACS/RIS search and a web system for searching images of the biomedical literature is the goal of this paper. Objectives are to identify similarities and differences in search behaviour of the two systems, which could then be used to optimize existing systems and build new search engines.

Log files of the ARRS GoldMiner medical image search engine (freely accessible on the Internet) containing 222,005 queries, and log files of Stanford's internal PACS/RIS search called radTF containing 18,068 queries were analysed. Each query was preprocessed and all query terms were mapped to the RadLex (Radiology Lexicon) terminology, a comprehensive lexicon of radiology terms created and maintained by the Radiological Society of North America, so the semantic content in the queries and the links between terms could be analysed, and synonyms for the same concept could be detected. RadLex was mainly created for the use in radiology reports, to aid structured reporting and the preparation of educational material (Lanlotz, 2006) [1]. In standard medical vocabularies such as MeSH (Medical Subject Headings) and UMLS (Unified Medical Language System) specific terms of radiology are often underrepresented, therefore RadLex was considered to be the best option for this task.

The results show a surprising similarity between the usage behaviour in the two systems, but several subtle differences can also be noted. The average number of terms per query is 2.21 for GoldMiner and 2.07 for radTF, the used axes of RadLex (anatomy, pathology, findings, ...) have almost the same distribution with clinical findings being the most frequent and the anatomical entity the second; also, combinations of RadLex axes are extremely similar between the two systems. Differences include a longer length of the sessions in radTF than in GoldMiner (3.4 and 1.9 queries per session on average). Several frequent search terms overlap but some strong differences exist in the details. In radTF the term "normal" is frequent, whereas in GoldMiner it is not. This makes intuitive sense, as in the literature normal cases are rarely described whereas in clinical work the comparison with normal cases is often a first step.

The general similarity in many points is likely due to the fact that users of the two systems are influenced by their daily behaviour in using standard web search engines and follow this behaviour in their professional search. This means that many results and insights gained from standard web search can likely be transferred to more specialized search systems. Still, specialized log files can be used to find out more on reformulations and detailed strategies of users to find the right content.

* Corresponding author.
E-mail address: mdeartea@andrew.cmu.edu (M. De-Arteaga).

http://dx.doi.org/10.1016/j.jbi.2015.04.013
1532-0464/

1. Introduction

The rate at which new medical information and research are produced is rapidly increasing [2]. Right now, it is already impossible for a physician to keep up to date on the material produced in his speciality, since the volume of publications surpasses what can be reviewed [3]. Therefore, search engines gain importance in the medical field by providing a way of finding relevant information when needed in a limited time. Moreover, when a physician searches for information, the ultimate goal is often to make a decision, such as whether to order a test or prescribe an additional imaging exam [4]. The fact that the results delivered are used as decision criterion, makes validity and relevance essential for medical search engines. The growth in the amount of information has a particularly rapid increase for imaging, which is estimated to occupy 30% of world storage in 2010 [2], and radiologists need to make decisions on what to describe in the radiology report.

Nonetheless, search engine performance still fails to fully respond to physician's needs. The latter not only lack the time to keep up to date with research, they also lack the time needed to find what they need in the moment they make a decision. Often, they only have five minutes to look for the answer to a question [5]. In a study conducted by Ely et al. [6], physicians were unable to find an answer to over 40% of a set of clinical questions. As part of their conclusions, the authors suggest physicians should change their search strategies. However, another way to tackle the problem is understanding how they search and changing the way information is retrieved accordingly, in order to fulfil the users' needs. For radiologists the information search behaviour is likely to be specific, with the information needs mainly being linked to images [7].

Log files provide valuable information to analyze user behaviour. They have been used to analyze how retrieval in medical search systems is conducted using MedLine [8,9] and GoldMiner [10–12]. In general, log files from this type of web search engines, as well as from Google and Bing, have been widely used in order to understand user behaviour [13,14] and improve search engines. However, both GoldMiner[1] and PubMed[2] are web search engines for content of the biomedical literature. Research based on log files from closed domain search systems, particularly in the hospital field, is yet to be done. This paper attempts to take one step into this direction, analyzing the query log files from the Stanford radiology search engine radTF. These files have never been used for this type of research, and particularly the comparison with a similar image search system on the web (GoldMiner) is interesting, since it provides an opportunity to analyze similarities and differences between the two. A screenshot of the GoldMiner interface with the query *ACL tear* can be seen in Fig. 1.

Tsikrika et al. [10] focused on how users formulate and reformulate queries, and Rubin et al. [11] attempt to understand what users look for by mapping queries to the RadLex[3] (Radiology Lexicon) terminology. De-Arteaga et al. [12] combine both, using a larger and more complete set of GoldMiner logfiles. In this paper, a similar analysis is done, this time for radTF search log files, and results are compared to those from GoldMiner, since one of the main questions driving this research is: how does user behaviour differ between search on the web and search in a clinical scenario (RIS/PACS)? In this case both search systems allow text search in metadata to search for images. Content-based image retrieval [15] was not a main target of this research, even though some results can likely be generalized to this domain as well. Content-based retrieval alone has often had limited performance [16] but mapping key words to semantics and combining this with visual retrieval can allow for powerful ways to create queries and obtain focused results. There are some marked differences between the target systems, as radTF performs searches in unstructured radiology reports to find clinical cases, while GoldMiner searches in figure captions of the biomedical literature. Even though GoldMiner is a specialized search engine, it can be accessed by anyone on the Internet, while radTF is only available inside the hospital and can only be accessed by physicians who work there. GoldMiner does simple stemming and is based on boolean search – absence and presence of words. radTF allows semantic search (as shown in Fig. 2), meaning that terms are mapped to the RadLex ontology [1], synonyms are detected and mapped to the same term. Negations are also detected, which has a big impact on the results, as a large part of information in radiology reports is actually the absence of specific findings or patterns, so simple word presence would not deliver good results in this case. Word stems and wildcards can also be used to formulate queries that aim at finding several related terms.

Therefore, GoldMiner and radTF do not necessarily have the same users, and even when the users are the same they might be used in different circumstances: radTF is only available inside the hospital, therefore users (mainly radiologists) access it while working on clinical cases, research and teaching, whereas GoldMiner can be accessed from workplaces as well as from home or on the move. Additionally, GoldMiner provides access to images from peer-reviewed journals, while radTF accesses a Picture Archiving and Communication System (PACS) and radiology reports, therefore the images vary between the two. GoldMiner searches in figure captions for images that are in JPEG format, whereas radTF searches in radiology reports and images in DICOM (Digital Imaging and Communications) format, meaning it might retrieve full tomographic series.

Given there is a difference in both the content and the potential users of the search engines, comparing them can help to understand how these differences impact the browsing strategies and also the way the search engines needs to be built. In addition to this, relating radTF's user behaviour to that of Web search engine users provides information and means to interpret results, since the latter have been studied further. For example, determining similarities between user behaviour in the two systems could help identify findings that can be generalized in terms of strategies to support the user. For these reasons we believe that comparing two slightly different systems that search for medical images – mainly from the radiology domain – can help us gain more insight into how two groups of users access information and formulate queries, despite the apparent differences between the two systems.

Understanding user strategies and the way they vary according to the user, the situation the user is in and the features of the search engine are key to improve the image retrieval systems. Search engines have to be able to fulfil the users' needs, therefore they must not only deliver valid and relevant information but they have to be able to do it in the right time and in the right format so the user can take advantage of the information provided.

This paper in organized as follows: a description of the data, the preprocessing, and the methods used to describe search behaviour can be found in Section 2. Section 3 includes the results, with comparisons between the two search engines. These results and possible interpretations are discussed in Sections 4 and 5 contains the conclusions.

2. Methods

2.1. Data sources

ARRS GoldMiner (American Roentgen Ray Society) is a web-accessible medical image search engine developed by the

[1] http://goldminer.arrs.org/ accessed September 25, 2014.
[2] http://www.pubmed.gov/ accessed September 25, 2014.
[3] http://www.radlex.org/.

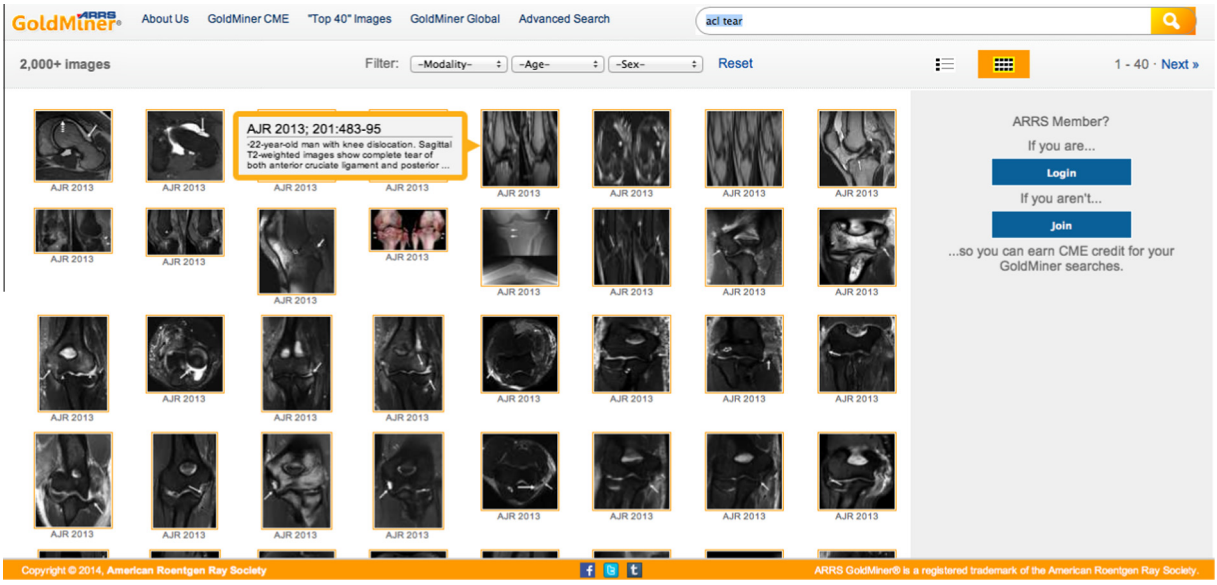

Fig. 1. Screenshot of the GoldMiner web interface with the query *ACL tear*.

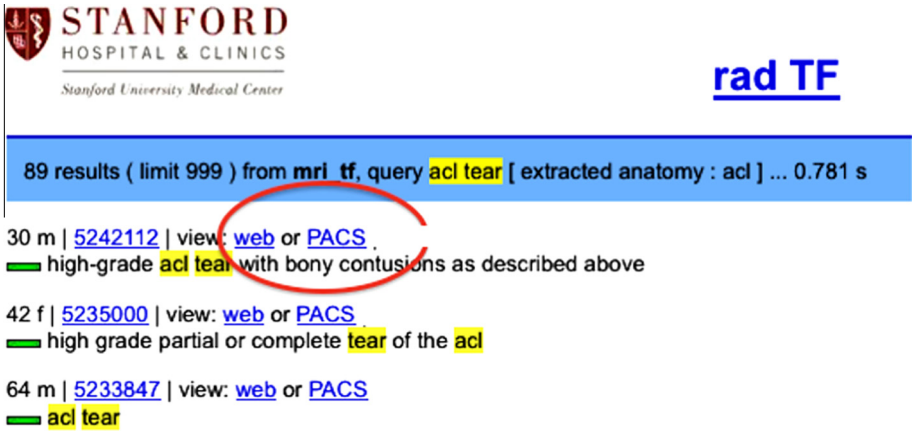

Fig. 2. Screenshot of the radTF interface that shows the semantic analysis and the possibility to view cases either in a web viewer or the PACS viewer.

ARRS [17]. It provides access to 420,000 selected images from peer-reviewed medical journals of selected journals that include the MeSH term *diagnostic imaging*. The system is openly accessible on the web and free of charge. In this paper, a log file containing 222,005 consecutive queries is analysed. Time stamps of each query are available, and identification of the searching person is possible via IP (Internet Protocol) addresses, which were mapped to consecutive numbers to preserve privacy. This allows session information to be analysed, even though it must be taken into account that several people could potentially use the same computer and thus the same IP address.

Unlike GoldMiner, which can be accessed by anyone on the Internet, radTF is only available for physicians who work at the hospital. It is mainly accessed from the shared computers used for the image reading and preparation of the radiology reports. RadTF provides access to 2 million internal cases of the Stanford radiology department, ranging from x-rays with a single image to MRIs where single series can contain up to 20,000 images. Cases include the images and radiology reports, and search is performed exclusively in the unstructured radiology reports. For this analysis 18,068 logs of consecutive queries were used. IP address and time stamp information are available as well. As the computers are shared the IP addresses of the same computer can be linked to several searchers. However, time stamps help establish sessions based on inactivity periods, which is particularly effective since a large number of computers is available for these tasks, so it is unlikely that many people share the same computer in a short period of time.

The preprocessing was done the same way it was done by Tsikrika et al. [10]. All special characters were removed and queries were lower-cased, selected medical imaging modalities were normalized (for example, "XR", "X-ray" and "xray" were associated to a single same term), and consecutive identical queries in the same session and with the same number of results were considered as a single query. The majority of these entries correspond to people browsing through several results pages of the same query. All empty queries are also removed. Much of the preprocessing criteria is based on manual analysis of a subset of the data.

2.2. Content-related attributes

2.2.1. RadLex mapping

In order to analyze the content of queries, aiming to determine what type of images users search for, mapping of the free text

queries to the RadLex ontology was used. As Rubin et al. [11] do, this radiology ontology from the Radiological Society of North America (RSNA) containing over 30,000 terms was used to gain insight into the semantic content of queries. The mapping was done by a system that, given a lexicon, automatically assigns biomedical categories to terms [18]. The system used was originally developed for UMLS and the Geneontology. It was not modified in this step and simply the RadLex ontology with its synonyms was used as input terminology. A small set of 100 examples per category was manually controlled to estimate the quality, which was considered very good with only a few uncorrect mappings, as detailed in the results section. Each term found in RadLex belongs to one of the following axes: Imaging protocol, Report, Procedure, RadLex descriptor, Property, Anatomical entity, Imaging observation, Process, Imaging modality, Non-anatomical substance, RadLex non-anatomical set, Report component, Procedure step, Object and Clinical finding, which are the main RadLex axes. In addition to the basic mapping, the type of mapping is also considered. A query might be entirely mapped to a single RadLex term, every single term of the query might be mapped to a term in the ontology, or a portion of the query might be mapped. Thus, there are three types of mapping – *exact*, *all terms* and *partial*. The mapping also allows for small linguistic changes in the terms such as shifted characters. There is an additional category – *no mapping* – for queries that cannot be mapped to RadLex by any of the three methods. Such cases can be explained by different reasons. A significant number of spelling mistakes occur in queries and even though we consider lexical similarity, spelling mistakes can still prevent mapping; some terms are not linked to the medical domain, and some correspond to physicians' personal names, when the user is searching reports of a specific radiologist. Finally, certain terms are simply missing in RadLex.

As a result, each query is tagged with one of the four types of mappings defined and is mapped to x RadLex axes, $0 \leqslant x \leqslant N$, where N is the number of terms in the query.

2.3. Formulation and reformulation of queries

For both sets of query log files, each entry contains the query itself, a timestamp, an IP address (mapped to consecutive numbers to keep anonymity) and the number of results obtained. The way users formulate queries was analysed by looking at the length of queries, the usage of common terms and the frequency of queries in the log files.

Additionally, based on the work done by Tsikrika et al. [10], but taking advantage of the availability of IP addresses that make it easier to determine when two queries were executed from the same machine, together with timestamps, query session analysis was performed. A session is defined as the set of consecutive queries with the same IP address, within less than 30 min of inactivity between them [19]. For each session, the type of reformulation was determined for consecutive queries: terms can be added, removed, single terms can be changed or the whole query might be reformulated. Other characteristics of the session such as session length measured in number of queries during a session are also studied.

3. Results

3.1. What do users search for?

Given the four different types of RadLex mappings, radTF has 16.4% (2,722) of its queries mapped exactly to a RadLex concept, 37.3% (6,212) of them partially mapped and 7.4% (1,224) having all the terms mapped, even though the query itself does not correspond to a RadLex term. The remaining 39% (6,482) were not mapped at all. For GoldMiner, these numbers are, respectively, 18.2% (36,372), 38.4% (76,928), 5.7% (11,419) and 37.8% (75,642). Fig. 3 shows these results. The high number of terms not mapped to a RadLex term was surprising for the case of radTF, since it is a radiology search engine and thus it could be expected it would have a higher percentage of matches to RadLex than GoldMiner. A very detailed analysis of the results of the RadLex mappings would go beyond this article but to verify the quality of the results in general the first 100 results in each of the categories were analysed for the radTF log files to estimate the quality of the mapping. As an outcome the *exact matches* were in 100% correct basically exactly the same term, same for the category of *all terms matched* where all of the first 100 were correct. This accounts for 23.8% with very high accuracy for radTF (including the exact matches and the matches of all terms). Each of the 100 analysed queries from the *partial match* category had at least one extra term in the query. These unmatched terms include abbreviations, empty words, or anatomy terms that are not included in the ontology. Even though sometimes unmatched terms seem irrelevant for the search, there are cases in which important anatomy terms are not identified (such as "myocard" or "vena cava"). The most interesting category is that of *unmatched queries*, where the reason for it not being matched can be traced back to a large variety of reasons; in the first 100 queries there were 2 numbers, 13 personal names, 11 non medical terms, 7 terms with spelling mistakes, 14 abbreviations, 17 queries where for us no meaningful term could be identified in the query, and in 17 cases the category was not clear to determine. The largest section are terms that are short forms of real terms, as the radTF system can work with wild cards to abbreviate search terms, which in the mapping system is not taken into account. Some of these points can be taken into account for optimizing mapping tools, and although this was not the goal of this article, results presented here could potentially be used for this purpose. The first 100 terms might not be fully representative, but we wanted to get an idea of the mapping quality, which despite some mistakes seems to perform very well.

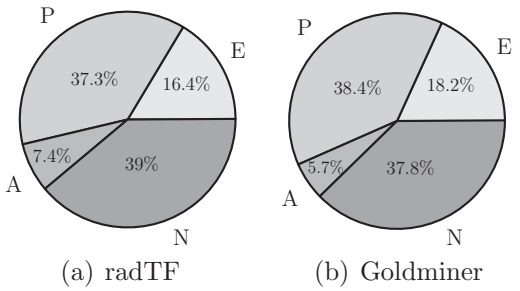

Fig. 3. Type of RadLex mapping in radTF and GoldMiner. E: Exact, A: All terms mapped, P: Partial, N: No Mapping.

Once the frequency and the way in which RadLex terms appear in the queries was observed, the next step is to inquire to which axes the RadLex terms belong to. In both cases, the most frequent axis is *clinical finding*; in radTF's case, it appears in 38.9% (6,467) of the queries, and in GoldMiner it can be found in 39.8% (79,721). The second most common axis in both search engines is *anatomical entity*, which occurs in 16.7% (2,784) of radTF's queries and 19.4% (38,791) of GoldMiner searches. The number of results a query has is closely related to the way in which the user combines terms. Therefore, the present study analyses the co-occurrence between Radlex axes. This information could later be used by the engine to give the user appropriate suggestions. Table 1 shows the number of queries in which two given axes co-occur in radTF's log files. Table 2 contains the same information for GoldMiner. The

Table 1
Co-occurrence of RadLex axes in radTF queries. CF: clinical findings, O: object, AE: anatomical entity, NS: non-anatomical substance, RD: RadLex descriptor, PP: property, P: procedure, PS: procedure step, IO: imaging observation, IM: imaging modality, RC: report component, R: report, PC: process.

	CF	O	AE	NS	RD	PP	P	PS	IO	IM	RC	R	PC
CF	6467	–	–	–	–	–	–	–	–	–	–	–	–
O	10	97	–	–	–	–	–	–	–	–	–	–	–
AE	834	30	2784	–	–	–	–	–	–	–	–	–	–
NS	40	0	18	230	–	–	–	–	–	–	–	–	–
RD	776	12	487	14	2192	–	–	–	–	–	–	–	–
PP	56	2	33	22	19	347	–	–	–	–	–	–	–
P	34	0	59	0	91	1	332	–	–	–	–	–	–
PS	10	0	2	3	2	3	1	94	–	–	–	–	–
IO	0	0	0	0	0	0	0	0	5	–	–	–	–
IM	1	3	2	1	1	0	0	0	0	13	–	–	–
RC	7	0	2	0	7	1	0	0	0	0	36	–	–
R	0	0	0	0	0	0	0	0	0	0	0	1	–
PC	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2
Co-occurrence of RadLex axes in GoldMiner queries. CF: clinical findings, O: object, AE: anatomical entity, NS: non-anatomical substance, RD: RadLex descriptor, PP: property, P: procedure, PS: procedure step, IO: imaging observation, IM: imaging modality, RC: report component, R: report, PC: process.

	CF	O	AE	NS	RD	PP	P	PS	IO	IM	RC	R	PC
CF	79721	–	–	–	–	–	–	–	–	–	–	–	–
O	175	1243	–	–	–	–	–	–	–	–	–	–	–
AE	11787	229	38791	–	–	–	–	–	–	–	–	–	–
NS	150	4	116	1161	–	–	–	–	–	–	–	–	–
RD	8272	89	5217	55	22321	–	–	–	–	–	–	–	–
PP	225	7	166	7	189	1092	–	–	–	–	–	–	–
P	280	18	357	4	163	16	1889	–	–	–	–	–	–
PS	0	1	12	0	1	0	1	101	–	–	–	–	–
IO	97	6	488	2	543	16	11	0	4044	–	–	–	–
IM	552	25	580	2	249	9	23	0	12	2211	–	–	–
RC	2	0	5	0	3	0	0	0	0	0	10	–	–
R	4	0	1	0	0	0	1	0	0	0	0	16	–
PC	1	1	5	0	0	0	0	0	0	0	0	0	12

Table 3
Comparison of relative co-occurrence of RadLex axes in radTF and Goldminer queries. Positive values (green) show a higher relative co-occurence in radTF whereas negative values (red) depict a higher relative co-occurence in Goldminer. Color saturation increases with higher relative difference. CF: clinical findings, O: object, AE: anatomical entity, NS: non-anatomical substance, RD: RadLex descriptor, PP: property, P: procedure, PS: procedure step, IO: imaging observation, IM: imaging modality, RC: report component, R: report, PC: process.

	CF	O	AE	NS	RD	PP	P	PS	IO	IM	RC	R	PC
CF	-	-	-	-	-	-	-	-	-	-	-	-	-
O	0	-	-	-	-	-	-	-	-	-	-	-	-
AE	-0.068	0.004	-	-	-	-	-	-	-	-	-	-	-
NS	0.011	0	0	-	-	-	-	-	-	-	-	-	-
RD	0.026	0	0.016	0.004	-	-	-	-	-	-	-	-	-
PP	0.014	0	0.007	0.008	0	-	-	-	-	-	-	-	-
P	0.004	0	0.011	0	0.030	0	-	-	-	-	-	-	-
PS	0.004	0	0	0	0	0	0	-	-	-	-	-	-
IO	-0.003	0	-0.016	0	-0.018	0	0	0	-	-	-	-	-
IM	-0.018	0	-0.018	0	-0.008	0	0	0	0	-	-	-	-
RC	0	0	0	0	0	0	0	0	0	0	-	-	-
R	0	0	0	0	0	0	0	0	0	0	0	-	-
PC	0	0	0	0	0	0	0	0	0	0	0	0	-

diagonal, the co-occurrence of an axis with itself, simply shows the number of queries in which the axis can be found. To compare how this varies from one system to the other, Table 3 shows the difference between the relative co-occurrence of RadLex axes in radTF and Goldminer queries. Positive values (green) indicate the given co-occurence is more frequent in radTF whereas negative values (red) depict the correlation is more characteristic of Goldminer's users behaviour. RadTF has a higher occurence of *anatomical entity*

and *clinical finding* together, as well as slightly higher values for *modality* and *imaging observation* combined with *clinical finding* and *anatomical entity*. Goldminer has a slightly higher co-occurrence of *procedure* and *property* but at a rather low absolute level.

For both search engines, no query can be mapped to more than four axes at the same time. In radTF's case, 47.2% (7,859) are mapped to a single axis, 13% (2,161) are mapped to two axes, 0.8% (135) to three and only 0.01% (3) to four different axes. For GoldMiner, these percentages are, respectively, 49.4% (99,060), 11.7% (23,477), 1.1% (2,130) and 0.03% (52). The two systems have thus extremely similar characteristics in this respect.

3.2. How do users formulate queries?

The GoldMiner log file contains 222,005 queries, which are reduced to 200,361 after preprocessing. radTF's log files are composed of 18,068 queries, and 16,641 remain after preprocessing. For GoldMiner, 75,118 (37.5%) queries are unique, with each query having an average number of occurrences of 2.16. In radTF's case, 47.9% (7,965) of the queries are unique, and each query appears on average 1.6 times. If attention is drawn to the most frequently occurring queries, however, the ten most frequent radTF queries represent 3.2% (536) of all the queries, while for GoldMiner this proportion is 2.2% (4,504). Table 4 shows the ten most frequently occurring queries of each search engine. Understanding exact queries as previously defined, where an exact query is one that matches exactly a RadLex term, among the 50 most common exact queries of the two search engines, 12% (6) are common to both. This number rises to 56% (28) when analyzing the 50 most common terms. Again, this shows a similarity between the systems, even though there are many differences in the details. The total number of queries of the two log files also varies by a factor of more than ten, which can explain some of these differences. Even if the most frequent queries are not exactly the same, the fact that terms are similar between the two underlines a similarity in the behaviour, as the probability of the same terms occurring by chance is extremely low.

For radTF, 43% (7,162) of the queries contain one of the 100 most frequent terms; for GoldMiner, 45.7% of the queries contain one of the top 100 most frequent terms. Table 5 shows the ten most common terms for radTF and GoldMiner. The term occurrences are statistically different between the two with a binomial test. Still, they are in a similar range.

The average number of terms in a query is of 2.21 for GoldMiner and 2.07 for radTF. A graph showing the number of queries given the number of terms in the query can be seen in Fig. 4. For GoldMiner, 90.8% (15,621) of the queries are composed of three or fewer terms, while this type of query represents 93.9% of radTF queries. For counting the number of terms no stop words were removed. They were removed for the list of the most frequent terms, though, to focus on terms that carry information. For counting the word occurrences, no stemming was used as we were aiming at exact wordings of the queries. This obviously creates a higher number of different queries as there is less harmonization.

Table 4
Most common queries.

radTF	GoldMiner
ACL tear (1.26%)	Mega cisterna magna (0.41%)
Study (0.31%)	Baastrup disease (0.40%)
Appendicitis (0.30%)	Limbus vertebra (0.23%)
ACL graft tear (0.24%)	Toxic (0.21%)
Hepatic adenoma (0.22%)	Cystitis cystica (0.20%)
Annular pancreas (0.20%)	Buford complex (0.14%)
Varicocele (0.19%)	Thornwaldt cyst (0.14%)
Perthe (0.18%)	Splenic hemangioma (0.13%)
Chiari (0.17%)	Double duct sign (0.12%)
Angiosarcoma (0.15%)	Cystitis glandularis (0.12%)

Table 5
Most common terms.

radTF	GoldMiner
Tear (2.62%)	Cyst (3.17%)
Fracture (2.03%)	MRI (1.89%)
ACL (1.81%)	Disease(1.76%)
Cyst (1.55%)	CT (1.75%)
Normal (1.36%)	Fracture (1.68%)
Liver (1.35%)	Tumour (1.61%)
Mass (1.19%)	Syndrome(1.49%)
Renal (1.12%)	Liver (1.24%)
Hepatic (1.11%)	Pulmonary (1.21%)
Carcinoma (0.97%)	Sign (1.14%)

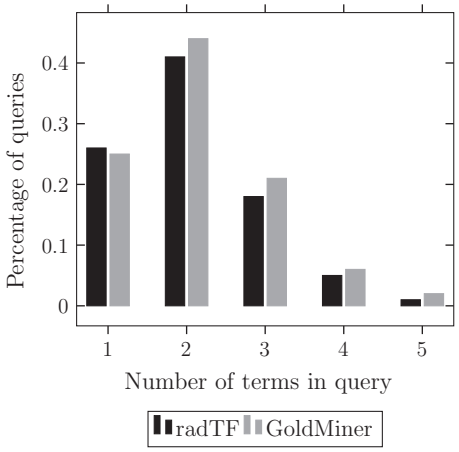

Fig. 4. Ratio of queries and the number of terms in the query.

3.3. How often and in which way do users reformulate queries?

Defining a session as the group of queries coming from the same IP address within a timespan of a maximum of 30 min, the 200,361 GoldMiner queries can be grouped into 103,029 sessions, and the 16,641 radTF queries correspond to 4,870 sessions. While the first search engine has an average number of queries per session of 1.9, the second one has an average of 3.4 queries in each session.

This could be due to the fact that radTF users search exhaustively until they find what they are looking for. As radTF is often used for preparing radiology conferences, this can be due to long sessions when the search goal is to find a large spectrum of cases, which can require many queries and a longer time. A few very long sessions in radTF that do not occur in GoldMiner underline this hypothesis. It could also be potentially caused by a better performance of GoldMiner's search engine, being faster at delivering what the user is looking for. It might also be due to users being satisfied much more easily, as it is more likely that GoldMiner is used for browsing with no particular goal in mind. An alternative explanation could be linked to the fact that in radTF computers are shared, in which case this might not be an optimal way to define sessions, since queries of more than one user can be combined into a single session. This last scenario is unlikely, though, as there is a much larger number of computers than there are users. It is expected that rarely a computer will be occupied by different people in a short period of time.

The fact that radTF is used on shared computers is reflected in the difference between the average number of sessions per user: for GoldMiner it is 2.3 while Stanford has a mean of 4.8.

Going back to the number of queries in a session, the percentage of sessions with two queries in GoldMiner is 16.9% (17,379) and 18.5% (899) in radTF. In the case of three queries per session, these percentages are of 8.2% (8,453) and 11.6% (563), respectively. A big difference, however, can be observed in the extremes. The number of sessions with only one query is of 62.8% (64,679) for GoldMiner, while for radTF it is only 40.2% (1,956). At the same time, while for GoldMiner only 11.4% (11,712) of the sessions contain more than four queries, for radTF this number rises to 29.8% (1,452). A hypothesis to why this happens is related to the fact that radTF is often used for very detailed search sessions, such as that required when preparing a collection of cases for a conference.

In order to understand what type of reformulations users do, three types of reformulations where defined: *generalization* – terms are removed, *specification* – terms are added, and *reformulation* – the query is modified by adding and removing terms.

Understanding a query pair as a pair of consecutive queries in which the second one is a modification of the first one, in GoldMiner's case, among the 97,315 query pairs, 13.5% (13,139) are specifications, 17.2% (16,757) are generalizations and 31.4% (30,622) are modifications. In radTF's case, 10.2% (1,198) correspond to specifications, 9.9% (1,169) are generalizations and 26.7% (3,138) are reformulations. The remaining 37.1% (36,056) and 43.5% (5,116) of GoldMiner's and radTF queries, respectively, correspond to queries that do not share any terms with the prior query.

4. Discussion

The way in which users search for medical images of the literature in GoldMiner and for PACS images in radTF was analysed and both behaviours were compared. As it was noted in the introduction, both search engines differ in the users, the circumstances in which people access them, the techniques for ranking the results lists and the type of images each system contains. Commonly, people search the former to find example cases of disorders often for research or teaching. People search in the latter for decision support, to look for studies related to diseases of a case currently diagnosed, or for a clinical research that requires certain cases to be found.

Therefore, broad differences between browsing strategies were expected. However, the analysis in this paper shows that the content of the queries and the way in which they are formulated and then reformulated in a query session are surprisingly similar, even though some notable differences occur; also, the frequent RadLex axes in both systems are almost identical. This common behaviour might be explained by the fact that today people are used to searching with Google (or very similar search engines such as Bing) on a daily basis, and they perform any other search they do in the same or a very comparable way, regardless of whether the search engine or the domain are different [7].

There are some differences, though, that merit a more detailed analysis. In the most frequent queries there is only a small overlap between the two systems, but this can be due to the fact that the log files cover only a limited period and the number of repetitions of queries is generally quite low with 50% of the queries occurring only a single time. Some terms are characteristic, though, as in radTF the term 'normal' is frequent, which can be expected in a clinical system where the first comparison is often with a normal case, whereas the images of the literature in GoldMiner will likely contain almost no normal cases. Search for radiologist names are also common in radTF, even though it is not reflected among the most frequent terms. In GoldMiner search, figure captions in the literature rarely contains personal names.

Another aspect in which a major difference was found is the number of queries per session (1.9 for GoldMiner, 3.4 for radTF). This might be caused partly by the fact that radTF is accessed from shared computers, therefore queries from different users might be mixed up in one single session or attributed to a single user. A plausible explanation for the large number of single query sessions in GoldMiner, may be linked to people testing the web page without a clear goal, while the number of really long sessions in radTF can be linked to people preparing case collections for conferences and research, thus running many queries in a row. The number of sessions per user (2.3 for GoldMiner, 4.8 for radTF) is even more likely to be linked to shared computers as the current setup does not allow us to identify single users reliably. radTF sessions appear to have more queries that do not share any terms with the prior query than GoldMiner, which can also be linked to people using shared computers, or the same person diagnosing several cases on the same computer and thus changing topic within the same session. In order to understand whether this is the case or, indeed, radTF users spend more time searching, an alternative way of defining sessions would be needed. Determining whether two queries are done by the same user is not easy in this scenario, given all queries are on the same domain, and sometimes they might be reformulated in such a way that two queries do not have any common words even though they do come from the same person and are on the same subject (for example, "hepatic" is sometimes replaced by "liver"). This would make it hard to define sessions based on the query terms.

A limitation of this work is that the characteristics of the search system are not taken into account. radTf uses negation detection, a more complex semantic analysis, and provides a ranking of the results, which will likely influence user behaviour and quality of results, as the absence of a concept or term such as smoker in a text is not the same as a negated statement, which frequently occurs in radiology reports and needs to be taken into account. Still, such a difference in results should also have a major influence on the way that people search for images, thus the analysis in this paper shows that there are many more similarities than differences in this behaviour.

5. Conclusions

Log files of how users formulate information needs and how they interact with query systems have been used in many ways in the past, particularly to understand the user behaviour and build better information retrieval systems. This article focuses on a comparison between two systems that allow search for medical images. The two systems vary strongly in the indexed content and the potential users. GoldMiner is a search system for images in the biomedical literature and is available on the Internet. radTF, on the other hand, is an internal hospital system that allows search in the radiology reports in the RIS and related images stored in the PACS. radTF is thus mainly used by physicians that are diagnosing cases or potentially have other tasks related to teaching and research inside a hospital. The GoldMiner users are potentially much more heterogeneous as the system is openly accessible on the Internet and can also be used by people without medical knowledge. Still, we can assume that a majority are users related to radiology who are looking for specific images related to their daily tasks, such as illustrating presentations or finding reference images and related articles for diagnosis. GoldMiner uses boolean search including extracted RadLex concepts, whereas radTF maps query terms to RadLex concepts and also detects negation, which is particularly important in radiology reports that contain in large part the absence of concepts. The fact that a patient is a

non-smoker is very different from the term smoker or smoking appearing in a text.

Despite the differences in indexed content and the potential users, the analysis of the results shows that the search behaviour is very similar. The number of terms in a query, the types of RadLex axes used and the types of reformulations are all extremely comparable. The only differences are the session length, which is on average longer for radTF than GoldMiner, and there are also differences on the most frequent queries, although they share several common terms.

Based on the comparisons we can clearly say that there are quite a few similarities in the ways people search clinical records and how they search on the web for medical images. This can be used to improve existing query systems as there is a broad body of knowledge for more general web search and this likely applies in a very similar form for search in clinical records. The reformulations of a query in a session can mean that the results were not satisfying, and the way users reformulate a query can be used as suggestions for future users, to improve their user experience. These can also be used to find common spelling mistakes. It would be extremely interesting to know which queries a user considers successful. This can sometimes be derived from click data of documents or images that the query retrieves. This information was unfortunately not available for radTF nor GoldMiner. However, for GoldMiner we have the number of results of a query, so we can assume that queries with zero results are not successful. Queries with too many results might not be useful either, since they might not be specific enough.

When manually analyzing the query terms that seem problematic we can find an important number of misspelled queries that are corrected by the user in the following query. A good retrieval system can potentially do this automatically. A few queries were also found in languages other than English. In both systems this would lead to bad query results, as both index English documents. Wikipedia can potentially help with this as it is possible to find the language of terms via their interface. Already knowing whether a query is medical or not (GoldMiner had queries such as "Happy New Year") can help a retrieval system respond to the user without delivering potentially non-relevant results.

We can also envision the use of these outcomes for improving RadLex, adding terms that are queried regularly but cannot be mapped to the ontology.

This article analysed image search behaviour in two quite different systems and scenarios. The results can be used to rethink medical image search and many of the problems identified can likely be solved easily, leading to better medical image search.

Acknowledgement

This work was partly support by the European Union in the KHRESMOI project (grant agreement 257528).

References

[1] C.P. Lanlotz, RadLex: a new method for indexing online educational materials, Radiographics 26 (2006) 1595–1597.
[2] Unknown, Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data, Submission to the European Commission, 2010. <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>.
[3] A.G. Fraser, F.D. Dunstan, On the impossibility of being expert, BMJ 341 (2010).
[4] W. Hersh, Information Retrieval – A Health and Biomedical Perspective, second ed., Springer, 2003.
[5] A. Hoogendam, A.F. Stalenhoefand, P.d.V.F. Robbé, A.J. Overbeke, Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed, J. Med. Internet Res. 10 (4) (2008).
[6] J.W. Ely, J.A. Osheroff, S.M. Maviglia, M.E. Rosenbaum, Patient-care questions that physicians are unable to answer, J. Am. Med. Inform. Assoc. 14 (4) (2007) 407–414.
[7] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, S. Kriewel, H. Müller, A survey on visual information search behavior and requirements of radiologists, Methods Inform. Med. 51 (6) (2012) 539–548.
[8] J.R. Herskovic, L.Y. Tanaka, W. Hersh, E.V. Bernstam, A day in the life of PubMed: analysis of a typical days query log, J. Am. Med. Inform. Assoc. 14 (2) (2007) 212–220.
[9] R.I. Dogan, G.C. Murray, A. Névéol, Z. Lu, Understanding PubMed® user search behavior through log analysis, Database 2009 (2009) bap018.
[10] T. Tsikrika, H. Müller, C.E. Kahn Jr., Log analysis to understand medical professionals' image searching behaviour, in: Proceedings of the 24th European Medical Informatics Conference, MIE'2012, 2012.
[11] D.L. Rubin, A. Flanders, W. Kim, K.M. Siddiqui, C.E. Kahn Jr, Ontology-assisted analysis of web queries to determine the knowledge radiologists seek, J. Digital Imag. 24 (1) (2011) 160–164.
[12] M. De-Arteaga, I. Eggel, C.E. Kahn Jr, H. Müller, Analyzing medical image search behaviour: semantics and prediction of query results, J. Digital Imag. (2015) 1–10.
[13] L.D. Catledge, J.E. Pitkow, Characterizing browsing strategies in the World-Wide Web, Comput. Netw. ISDN Syst. 27 (6) (1995) 1065–1073.
[14] B.J. Jansen, D.L. Booth, A. Spink, Determining the user intent of web search engine queries, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 1149–1150.
[15] H. Müller, N. Michoux, D. Bandon, A. Geissbuhler, A review of content-based image retrieval systems in medicine-clinical benefits and future directions, Int. J. Med. Inform. 73 (1) (2004) 1–23.
[16] A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, in: Working Notes of CLEF 2013 (Cross Language Evaluation Forum), 2013.
[17] C.E. Kahn Jr., C. Thao, GoldMiner: a radiology image search engine, Am. J. Roentgenol. 188 (2008) 1475–1478.
[18] P. Ruch, Automatic assignment of biomedical categories: toward a generic approach, Bioinformatics 22 (6) (2006) 658–664.
[19] R. Jones, K.L. Klinkner, Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM, 2008, pp. 699–708.